DOCUMENT RESUME

ED 427 084 TM 029 467

AUTHOR Burdenski, Thomas
TITLE A Review of the Latest Literature on Whether Statistical Significance Tests Should Be Banned.
PUB DATE 1999-01-00
NOTE 35p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 21-23, 1999).
PUB TYPE Information Analyses (070) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Effect Size; Hypothesis Testing; Literature Reviews; Social Science Research; *Statistical Significance
IDENTIFIERS *Null Hypothesis; Research Replication

ABSTRACT

Controversy over the merits of Null Hypothesis Statistical Significance Testing (NHST) as a tool for advancing knowledge in the social sciences has intensified in recent years. Literature for and against the use of statistical significant tests is reviewed and three major limitations of these tests are summarized. The first is that "p" values themselves cannot be used as indices of effect size. A second limitation is the recognition that unlikely results are not necessarily interesting or important. A third limitation is that "p" values do not bear on the important issue of result replicability because statistical tests do not test the possibility that sample results occur in the population. A summary is also presented of what NHST can and cannot do. (Contains 1 table and 48 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 427 084

burdeski.wp1 1/12/99

Running Head: STATISTICAL SIGNIFICANCE TESTS

A Review of the Latest Literature on Whether Statistical Significance Tests Should Be Banned

Thomas Burdenski

Texas A & M University 77843-4225

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Thomas Burdenski

TM029467

Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, January 23, 1999.

BEST COPY AVAILABLE

Abstract

Controversy over the merits of Null Hypothesis Statistical Significance Testing (NHST) as a tool for advancing knowledge in the social sciences has intensified in recent years. The present paper reviews the literature concerning arguments both in favor of and opposed to the use of statistical significance tests and summarizes three major limitations of these tests. Finally, a summary is presented of what null hypothesis statistical significance tests can and cannot do.

Scientific controversy over the proper use of null hypothesis statistical significance testing (NHST) in the social sciences has smoldered for decades. In practice, researchers have long relied on the use of statistical significance tests without a clear understanding of what these tests can and cannot do. Empirical studies confirm that, indeed, many researchers do not understand what statistical significance tests do and do not do (cf. Nelson, Rosenthal & Rosnow, 1986; Oakes, 1986; Rosenthal & Gaito, 1963; Zuckerman, Hodgins, Zuckerman & Rosenthal, 1993). Similarly, content reviews of the most widely-used statistics textbooks show that even our most distinguished methodologists do not have a good grasp on the meaning of statistical significance tests (Carver, 1978).

NHST has flourished despite the fact that criticisms of Fisher's system of statistical induction date as far back as 1928 (Carver, 1978; Cronbach, 1975; Daniel, in press; McLean & Ernest, in press; Meehl, 1978; Morrison & Henkel, 1970; Neyman & Pearson, 1928; Nix & Barnette, in press; Oakes, 1986; Rozeboom, 1960; Thompson, 1993, 1998a, 1998b, 1998c, in press-a, in press-b, in press-c). A series of articles on these issues appeared in recent editions of the American Psychologist (e.g., Cohen, 1990; Kupfersmid, 1988; Rosnow & Rosenthal, 1989). Especially noteworthy are recent articles by Cohen (1994), Kirk (1996), Schmidt (1996) and Thompson (1996). Another signal of growing uneasiness about the pervasive misuse of NHST is a recent decision by the APA Board of Scientific Affairs to launch a Task Force on Statistical Inference (Azar, 1997; Shea, 1997).

NHST: Arguments Pro and Con

Views on Null Hypothesis Statistical Significance Testing in the recent literature can be arranged along a continuum ranging from those who defend its use (cf. Abelson, 1997; Cortina & Dunlap, 1997; Frick, 1996; Hagen, 1997; Rindskopf, 1997) to those who believe NHST should be banned (cf. Carver, 1978, 1993; Hunter, 1997; Schmidt, 1996; Schmidt & Hunter, 1997). Robinson and Levin (1997) and Levin (1998) take a more moderate view, but are basically test advocates. Kirk (1996) takes a moderate view, but emphasizes the importance of effect sizes. Cohen (1990, 1994) and Thompson (1993, 1996, 1998a, 1998b, 1998c, in press-a, in press-b, in press-c) can be best characterized as viewing NHST as a relatively unimportant tool for social science research, but one that must be used properly, and especially as emphasizing effect sizes and evidence of result replicability. Some of the defenses of NHST have been thoughtful, while others are seriously flawed (see Thompson, 1998b).

Among the detractors of NHST, Schmidt (1996) takes the hardest line:

> My conclusion is that we must abandon the statistical significance test. In our graduate programs we must teach that for analysis of data from individual studies, the appropriate statistics are point estimates of effect sizes and confidence intervals around these point estimates. We must teach that for analysis of data from multiple studies, the appropriate method is meta-analysis.

(p. 116)

Schmidt asserted that for nearly 50 years, reliance on NHST to interpret research data has led to serious misinterpretations and erroneous conclusions that have substantially impeded the advancement of knowledge in the social sciences. He contended that this misguided alliance has been based on three fundamental false beliefs.

First, many researchers have falsely believed that statistical significance indicates the probability of successful replications of a study. The second false belief is that statistical significance provides a measure of the importance or size of a difference or a relationship. The third false belief is that if there is no statistical significance in a test of difference or relationship, then the difference or relationship between variables is zero or so close to zero that it may be considered zero. He argued that this last belief has been most devastating to the research enterprise because it has led to the erroneous assumption that if the null hypothesis is not rejected, then it is accepted, and that the NHST determines whether or not a difference or relationship is real or probably occurred by chance. Schmidt (1996) issued the following challenge to supporters of NHST: "Can you articulate even one legitimate contribution that significance testing has made ...to the development of scientific knowledge? I believe you will not be able to do so" (p. 116).

In response to this challenge, Abelson (1997) argued that the generation of categorical statements by NHST, despite their provisional and uncertain status, has important benefits to the

development of scientific thought. Knowledge can grow by comparing results across studies from different times and places (as in meta-analysis) and it can also grow through the social process that ensues when findings are published, discussed and reacted to in discourse between researchers. Disagreement and controversy provide fertile soil for dissolving entrenched thinking and allow the cross-fertilization of ideas and the generation of new ideas for further research. Important new findings are what Abelson (1994) called the "lore of the field." The lore is informal and includes findings that do not always hold up in subsequent research, but the lore is qualitatively rich and includes procedural details upon which future investigators can base research. If a study with surprising results makes a theoretical contribution to the field, it is even more likely to become part of the lore. Inconsistencies in the lore lead to examination of the record and redrawn conclusions lead to revision of the lore. Thus, the categorical nature of NHST helps advance science because it provides a stimulus to which researchers can respond.

Frick (1996) contended that NHST is the optimal procedure for demonstrating sufficient empirical evidence to support an ordinal claim. He defined an ordinal claim as "one that does not specify the size of effect; alternatively, it could be defined as a claim that specifies only the order of conditions, the order of effects, or the direction of correlation" (p. 380). He distinguished ordinal claims from quantitative claims, which report a measure of effect size. He believes that ordinal claims are common in psychological research today and that the field of psychology may always have

laws and theories making ordinal predictions. In such a paradigm, he believes that NHST is an appropriate procedure. Similarly, experimenters frequently use ordinal laws for the prediction of ordinal theory or to test an ordinal law. Findings from such research are categorized as acceptable or unacceptable to enter the body of knowledge in psychology and NHST is an appropriate tool for making these determinations.

Similar to Abelson (1994), Frick (1996) saw value in the categorical nature of NHST because if there were only a handful of claims in any given area of psychology, it would be possible to assign them probabilities and then update those probabilities as new research is reported. Since there are hundreds and perhaps thousands of such claims, NHST provides a criterion for entrance into the corpus of knowledge that is considered established. Frick likened this function to the baseball hall of fame in which a ballplayer must receive 75% of the votes of sportswriters to be admitted--a player either receives enough votes to be elected or he does not.

Cortina and Dunlap (1997) agreed with Cohen (1994) and others who believe that statistical significance testing is abused, that interpretation of p values as the probability of the null hypothesis given the data is erroneous, that confidence intervals and effect sizes should be reported, and that the application of Modus Tollens to probabilistic statements can lead to problems. They disagreed with Cohen on four points.

First, they argued that the purpose of data analysis is to provide evidence about the strength of corroboration for the answer

to the research question based on theory, usually in the form of disconfirmation of alternative hypotheses. NHST, then, is a useful tool for ruling out hypotheses related to the null because in their opinion, no other analytic procedure is as effective as NHST for addressing three inter-related requirements for empirical corroboration: objectivity, exclusion of other alternate hypotheses, and exclusion of alternate explanations (confounds, sampling error, etc.).

Second, they argued that attacks on the logic of NHST are based on misleading examples, a misunderstanding of key concepts and faulty premises. Cohen (1994) gave the following example as a demonstration of the problem with using Modus Tollens form of logic with probabilistic statements:

> If a person is an American, then that person is probably not a member of Congress.
> This person is a member of Congress, therefore,
> This [sic] person is probably not an American. (p. 998)

Cortina and Dunlap (1997) contended that this example is problematic for two reasons. First, the consequent of the second half of the first statement is true in and of itself--any given person is probably not a member of the U.S. Congress. Consequently, almost any statement could be used as the first part of the first statement and that premise would not effect the veracity of the second half of the statement. Second, while the second half of the first statement stands alone ("...that person is probably not a member of Congress"), it is also true that being an American is a

necessary condition for becoming a member of the U.S. Congress. So they argued:

> In other words, while it is true that "If a person is an American, then that person is probably not a member of Congress," it is also true that if a person is a member of Congress, then that person has to be an American. It is because of these two aspects of the particular example chosen that the Modus Tollens breaks down. (Cortina & Dunlap, 1997, p. 166)

They contended that there are many cases in which the probabilistic use of Modus Tollens can be used to produce approximate probabilistic statements about hypotheses. For example, they cited the following as more representative of psychology than Cohen's statements about Congress:

> If Sample A were from some specified population of "normals," the Sample A probably would not be 50% schizophrenia.
>
> Sample A comprises 50% schizophrenic individuals; therefore,
>
> Sample A is probably not from the "normal" distribution. (p. 166)

Third, they also contended that a clear understanding of error rates make p values useful, regardless of the actual nature of a given population. Cortina and Dunlap (1997) asserted that Schmidt (1996) and Cohen (1994) falsely contended that Type I error rates are zero instead of .01 or .05 because the hypotheses of no effect

is never precisely true and it is never possible to falsely reject the null hypothesis. Their position is that perhaps the null is always false, but that this has nothing to do with the Type I error rate:

> The Type I error rate, α, is the probability that the null would be rejected if the null were true. Note that there is no suggestion here that the null is or is not true. The subjunctive *were* is used instead of is to denote the conditional nature of this probability. The Type I error rate is the probability that the hypothetical null distribution would produce an observed value with a certain extremeness... The .05 value is the Type I error rate, regardless of whether or not the null is true... Alpha is not the probability of making a Type I error. It is what the probability of making a Type I error would be if the null were true. One can, perhaps, argue that the term *Type I error rate* is misleading. (pp. 166-167)

Fourth, they Cortina and Dunlap asserted that the argument that NHST should be replaced by confidence intervals is absurd because the two are based on exactly the same information and both involve categorical decision-making of some form. They concluded that confidence intervals and power estimates should not be done instead of statistical significance tests, but rather, they should be done in conjunction with statistical tests.

Hagen (1997) took issue with Cohen's classic essay in the

American Psychologist (Cohen, 1994) and asserted that while there may be good reasons not to use NHST, Cohen's reasons are not among them. Specifically, he contradicted Cohen's conclusions that (a) the NHST does not tell us what we want to know; (b) the null hypothesis is always false; and (c) the NHST lacks logical integrity. In regard to NHST not telling researchers what they want to know, Hagen claimed that Cohen's example was flawed, not NHST. In Cohen's example the frequencies of schizophrenics and normal individuals in the population were 2% and 98%. Therefore, the probability of randomly drawing a normal individual is .98 and the probability of randomly drawing a schizophrenic individual is .02. Then using a Bayesian analysis, Cohen established a posterior probability of .60 which he referred to as "the probability that case is normal, given a positive test" (p. 999). Hagen asserted that Cohen erroneously implied that both .98 and .60 refer to the probability that the null hypothesis is true and that he defined the null hypothesis and alternative hypothesis in ways that NHST does not allow. Hagen added that Cohen's conclusion that NHST does not tell us what we want to know is only true when the researcher is seeking a frequency-based probability and his statement is false when we are satisfied with equating the null hypothesis with a subjective degree of belief or a confidence level.

Cohen (1994) made the following comment: "So if the null is always false, what's the big deal about rejecting it?" (p. 1000). Hagen (1997) interpreted this comment as a statement about "soft psychology," which he defined as referring to a study of variables from the same individual or entity or the study of differences

among intact groups. Under either of these conditions, Hagen believes that it is indeed true that the null hypothesis is almost always false. Hagen assumed Cohen's bold comment was a deliberate overstatement to "rattle us into more careful thinking" (p. 20) and in Hagen's experience with professional colleagues, Cohen's message has often been taken all too literally by researchers to erroneously mean that all null hypotheses are false under all conditions.

Hagen (1997) added that small differences will only be detected under the alternative hypothesis and not the null because when samples are drawn from the same population, the variance of absolute differences between samples becomes smaller as N gets larger. He stated:

> Type I error remains relatively constant no matter how large N becomes because the decreasing variance is reflected in the decreasing variance of the test statistic Thus, although it may appear that larger and larger Ns are chasing smaller and smaller differences, when the null is true, the variance of the test statistic, which is doing the chasing, is a function of the variance of the differences it is chasing. Thus, the "chaser" never gets any closer to the "chasee." (p. 20)

Hagen's (1997) third criticism of Cohen's assertion that the null hypothesis is always false centered around Cohen' belief that whenever groups are treated differently in any way, those differences will inevitably have a differential impact on the

groups. Hagen agreed that independent variables A and B will always produce differential effects on some variable or variables that can be measured theoretically, but he did not agree that A and B will always produce an effect on the dependent variable. A measurable impact on the dependent variable, naturally, is the only result that can lead to a rejection of the null hypothesis.

Hagen (1997) defended the integrity of NHST by pointing out that it does not have logical validity in the sense of formal logic because it is based on probability, but that does not mean that the procedure lacks practical utility. He also stated that certain forms of logical reasoning (e.g., Modus Ponens, Modus Tollens) may have formal logical validity, but may not be sound in practice. His example follows:

> If you contract AIDS, you will be healthy and happy.
>
> You did contract AIDS.
>
> You are healthy and happy. (p. 21)

In the AIDS example, Hagen argued that the argument is logical and valid, given the premise, but that logical validity has limitations for scientific argument. On the other hand, Hagen said that arguments can be defensible and reasonable even when they are not logically valid in the formal sense. In the following example (Cohen, 1994), the probabilistic argument is not logically valid because one could accept the premises but reject the conclusion. The argument, nonetheless, is based on defensible and reasonable data:

> If you contract AIDS, you will probably die of some opportunistic infection within ten years.

> You did contract AIDS.
>
> You will probably die of some opportunistic infection within ten years. (p. 22)

Hagen pointed out that most of the decisions we make throughout our lives are based on probabilistic premises, not on valid logic in the formal sense. He asserted that science has done well in the absence of arguments that are logically valid and that in the absence of an alternative, it will continue to do so through NHST, because nothing better is likely to come along. He stated:

> The logic of NHST is elegant, extraordinarily creative, and deeply embedded in our methods of statistical inference. It is unlikely that we will ever be able to divorce ourselves from that logic even if someday we decide to do so. (p. 22)

In response to Hagen (1997), Thompson (1998b) sidestepped the philosophical logical validity arguments raised by Hagen and focused on what he believed were omissions and three misinterpretations of Cohen in Hagen's article. Regarding omissions, he cited Cohen's (1994) criticism of nil versus non-nil hypothesis testing:

> Most researchers mindlessly test only nulls of no difference or of no relationship because most statistical packages only test such hypothesis. The use of what Cohen called nil hypotheses does not require researchers to thoughtfully extrapolate expected results from previous literature or theory. Instead, science becomes an automated, blind search

> for mindless, tabular asterisks using thoughtless hypotheses. (p. 799)

Thompson (1998b) criticized Hagen's failure to address Cohen's point that NHST would be more meaningful and useful if more thought was given to formulating meaningful hypotheses at the front end of the research process.

Thompson (1998b) referred to three apparent misrepresentations Hagen made when critiquing Cohen's article. First, Hagen argued that the null hypothesis makes a statement about the population. While psychologists want to know about the population to determine if the results will generalize and replicate, statistical tests do not provide that information. In that sense, argued Thompson, Cohen (1994) was correct when he said that statistical significance testing "does not tell us what we want to know" (p. 997).

Second, Thompson (1998b) argued that Hagen (1997) misrepresented Cohen's explanation as to why NHST are tautological (i.e., the null hypothesis is always false at some sample size). Thompson asserted that the null is always false in the sample "because the probability of any single point in a continuum of infinitely many sample statistics is itself infinitely small" (p. 799). And, because the null hypothesis is also never true in the population (although divergent views on this point can never be definitively resolved, because the population is infinite and unknowable), "if we fail to reject, it's only because we've been too lazy to drag in enough participants" (p. 799).

Third, Thompson (1998b) argued that Hagen (1997) was off target in his reasoning for recommending confidence intervals to

determine if the interval subsumes zero because using confidence intervals in this way invokes the same logic as NHST. On the other hand, if the confidence intervals in a study are examined in relation to the confidence intervals of previously conducted, related studies, the true population parameters will eventually be estimated across studies even if the parameter estimates are off in the first place. Thus, confidence intervals can be used effectively without invoking the flawed logic of NHST.

Thompson (1996, 1997) stated that just because researchers are inappropriately using and misinterpreting NHST does not mean that these tests ought to be abandoned. Instead, he made three recommendations. First, he urged the use of the phrase "statistically significant" instead of the phrase "significant" to reduce the tendency of researchers and consumers of research to infer that significance implies importance or has anything to do with importance.

Second, he recommended that effect sizes be reported in all studies, whether statistical tests are reported or not. This recommendation moves one step beyond the policy in APA publication manual's "encouragement" (APA, 1994, p. 18) to report effect sizes. Numerous empirical studies confirm that this vague encouragement has been utterly ineffective (cf. Keselman et al., in press; Kirk, 1996; Thompson & Snyder, 1998; Vacha-Haase & Nilsson, 1998).

Third, he recommended that researchers utilize strategies to determine result replicability, because statistical significance tests do not do so. In his view, because most researchers lack the stamina to conduct their studies more than once to evaluate true

"external" replicability, the second best option is to use "internal" replicability analyses using the jackknife, cross-validation and/or bootstrap approaches.

In a response to Thompson's recommendations (1996), Robinson and Levin (1997) agreed that the way authors report statistical results is a problem, but disagreed that rules should be formulated to mandate or regulate language. Because there is so much baggage and historical misuse, they recommended that the word "significant" be banished altogether and replaced with the phrase "statistically nonchance" or "statistically real." In reference to Thompson's (1996) recommendation that effect sizes should be routinely reported, they pointed out that while effect size provides valuable information about the magnitude of difference or relationship, they do not provide information about the probability that the estimated difference is due to chance (sampling error).

Robinson and Levin were concerned that allowing authors to promote "unusual" or "interesting" outcomes without evidence of probability would result in an onslaught of journal submissions fraught with chance or strange occurrences. Instead, they argued that journal editors ought to adopt a one-two editorial policy: first, require researchers to convince them that the research finding is not due to chance, then listen to the researcher's case for how impressive that finding is.

In reference to Thompson's recommendation that researchers conduct internal replicability analyses like jackknifing or bootstrapping, Robinson and Levin (1997) claimed that these techniques rely on combining participants in the current sample in

several different ways, with convergent statistical conclusions. Even when the internal replicability analyses are based on independent subsamples (i.e., cross-validation) they are still limited to the characteristics of the original sample and the original procedures followed by the experimenter. A major shortcoming of internal replicability, in their view, is that this technique does not take into account the biases and peculiarities associated with a one-time study based on a single sample. Thompson (1997) concurred, but noted that limited evidence of result replicability is superior to no evidence of result replicability, which is his view of statistical significance tests.

Kirk (1996) agreed with Thompson and others that NHST do not tell researchers what they want to know; that it is an exercise of dubious value since a decision to reject merely indicates that the research design had sufficient power to detect a true state of affairs (that the null is false), which may or may not be a large or useful effect; that NHST reduces a continuum of uncertainty to a dichotomous reject or fail-to-reject decision; and that the results of NHST are often misinterpreted. He applauded the efforts of Cohen (1969) and Glass (1976) for their pioneering work on measuring effect sizes in research designs. He particularly praised Cohen for developing the first effect size designated as such (Cohen's d) and for providing guidelines for interpreting the magnitude of d.

Kirk argued in favor of moving away from focusing on statistical significance to what he described as emphasizing "practical significance" determined by point estimates and

confidence intervals:

> The computation of a point estimate of the difference between A and B and a confidence interval for that difference requires no more information than a null hypothesis significance test. A confidence interval contains all of the information provided by a significance test and, in addition, provides a range of values within which the true difference is likely to lie. It is important to understand that a confidence interval is just as useful as a null hypothesis significance test for deciding whether chance or sampling variability is an unlikely explanation for an observed difference. Furthermore, a point estimate and confidence interval use the same unit of measurement as the data. This facilitates the interpretation of results and makes trivial effects harder to ignore. (p. 754)

According to Kirk (1996), when evaluating results using measurement scales that are familiar to the researcher, like IQ scales, a point estimate of a difference and a confidence interval could be used to decide whether results are trivial, useful or important. With measurement scales using units unfamiliar to the researcher, it is necessary to compute an effect magnitude and a confidence level for that effect magnitude and develop guidelines for deciding whether or not that magnitude of effect is of practical use.

Kirk (1996) argued that researchers should do all that they

reasonably can to supplement the use the NHST, but he also looked forward to the day that the NHST is phased out in textbooks, journal articles and instructional curriculum:

> The winds of change are about us. Many researchers share the belief that if our science is to progress as it should, we must get over our obsession with the null hypothesis significance tests and focus on the practical significance of our data. The appointment of the task force (by the APA Board of Scientific Affairs) may mark the beginning of a more enlightened approach to the interpretation of data. (p. 757)

Major Limitations of Statistical Significance Tests

Three major limitations of statistical tests have increasingly been recognized within the literature. First, as demonstrated by changes made in the Publication Manual of the American Psychological Association (APA), it has become evident that *p* values cannot themselves be used as indices of effect size:

> You can estimate the magnitude of the effect or the strength of the relationship with a number of measures that do not depend on sample size... You are encouraged to provide effect size information. (APA, 1994, p. 18)

These changes in APA editorial policy reflect a growing emphasis on reporting and evaluating effect size and analyzing the replicability of results observed in research and a movement away from statistical significance as an index of effect size. Thompson

(in press-a) noted,

> The calculated p values in a given study are a function of several study features, but are particularly influenced by the confounded, joint influence of study sample size and study effect sizes. Because p values are confounded indices, in theory 100 studies with varying sample sizes and 100 different effect sizes could each have the same single $P_{CALCULATED}$, and 100 studies with the same single effect size could each have 100 different values for $P_{CALCULATED}$.

Several elements contribute to the computation of $P_{CALCULATED}$: the sample statistics named in the null hypothesis (e.g., means, medians, standard deviations, Pearson r); the alpha or $P_{CRITICAL}$ value; and the sample size. For example, if a researcher wanted to compare the mean scores of males and females on an IQ test, and the sample mean for females was 115 and the sample mean for males was 110, in the classical NHST the researcher would assume these scores or "statistics" come from a population in which the two means are equal. The researcher must assume <u>something</u> about the population parameter means, because otherwise there would be an infinite number of answers to the question of what is the probability of the sample statistics for samples derived from the population. In practice, most researchers assume that the "nil" null exactly describes the population, because that is what most statistical packages assume (Thompson, 1998b).

Computations of $P_{CALCULATED}$ must also take into account the

sample size because sample statistics that do not exactly honor the null hypothesis are increasingly more unlikely as the sample size gets larger and larger. In other words, if a researcher had a sample size of 20 and the sample mean for men's IQ was 110 and the sample mean for women's IQ was 115, the probability of these sample statistics would become increasingly unlikely as the sample increases to an n of 40, 60, 80 and 100, because as sample size gets larger, "flukiness" or sampling error in the sample becomes increasingly unlikely. Conversely, as the n descends from 100 to 80, 60, 40 and 20, for a given set of sample statistics, the $P_{CALCULATED}$ gets larger and larger because "flukiness" or sampling error becomes more and more likely as the sample size decreases.

As Cohen (1990) pointed out, widespread use of Sir Ronald Fisher's invention of null hypothesis statistical significance testing emerged from the lure of a deterministic, mechanical research method that yielded clear-cut, yes-no decisions that ostensibly advanced scientific understanding through inductive inference by rejecting null hypotheses, usually at the .05 level. When the null hypothesis is rejected with an associated probability of less than .05 (say .02), it is erroneous to conclude that the probability that the null hypothesis is true is .02. This result does not inform the researcher about the truth of the null hypothesis, given the data. Rather, NHST tells the researcher the probability of the sample, presuming the truth of the null hypothesis, which is not the same thing (cf. Thompson, 1994).

Second, it has increasingly been recognized that unlikely results (i.e., results with a small p value) are not necessarily

interesting or important. Some highly improbable events, in fact, are completely inconsequential. For example, if one flips a silver dollar and it lands on its side, this result would be very unlikely; however, it is doubtful that the result would have particularly noteworthy effects on the coin, the coin flipper, or anyone else. Cohen (1994) piercingly portrayed the folly of naïve researchers who equate statistical significance with result importance:

> Because NHST p values have become the coin of the realm in much of psychology, they have served to inhibit its development as a science. Go build a quantitative science with p values! All psychologists know that *statistically significant* does not mean plain-English significant, but if one reads the literature, one often discovers that a finding reported in the Results section studded with asterisks become in the Discussion section highly significant or very highly significant, important, big! (p. 1001)

In valid deductive arguments conclusions cannot logically contain any information not also present in the argument's premises (Thompson & Snyder, 1998). So, as noted by Thompson (1993), "If the computer package did not ask you your values prior it its analysis, it could not have considered your value system in calculating p's, and so p's cannot be blithely be used to infer the value of research results" (p. 365). Thus, statistical significance tests cannot reasonably be used as an atavistic escape from

responsibility for defending result importance (Thompson, 1993), or to maintain a mantle of feigned objectivity (Thompson, in press-a).

Third, it has been increasingly recognized that p values do not bear upon the important issue of result replicability, because statistical tests do not test the possibility that sample results occur in the population. Instead, statistical significance tests assume that the null hypothesis exactly describes population values (e.g., parameter means, parameter correlation coefficients), and then evaluates the probability of the sample statistics, given the sample size and presuming that the sample(s) came from the assumed population (Cohen, 1994; Thompson, 1996).

Conclusions

Researchers would like to be able to draw inferences about the population from sample statistics because if they could legitimately do so, NHST would provide information about the replicability of results without having to undergo the arduous task of duplicating studies. If statistical significance did inform researchers about the population (which they do not), researchers would be able to predict with confidence that other researchers would be able to draw samples from the same population and identify the same relationships. If they could conduct a single study and know about the population, then they wouldn't need to repeat the same study over and over again to make sure their decisions are correct. Unfortunately, since the direction of the statistical inference is from the population to the sample, statistical significance testing does not tell researchers about replicability.

Of course, knowing the probability of the sample results that researchers already know is considerably less interesting than knowing the probable population parameters, which researchers do not know, but which they would like to know because knowing the population values would inform judgment regarding result replicability. Thus Cohen (1994) in his widely cited and influential article observed that the statistical significance test "does not tell us what we want to know, and we so much want to know that, out of our desperation, we nevertheless believe that it does!" (p. 997).

Furthermore, there is no point in learning about the probability of the sample because the "nil" null is never true in population anyway. According to Cohen (1990):

> The null hypothesis, taken literally (and that's the only way you can take it in formal hypothesis testing), is always false in the real world. It can only be true in the bowels of a computer processor running a Monte Carlo study (and even then a stray electron can make it false). If it is false, even to a tiny degree, it must be a case that a large enough sample will produce a significant result and lead to its rejection. So if the null is always false, what is the big deal about rejecting it? (p. 1308)

Notwithstanding the movement of the field away from the overemphasis on statistical significance, it remains important to understand the logic of these statistical tests. As Thompson (1996) noted:

We must understand the bad implicit logic of persons who misuse statistical tests if we are to have any hope of persuading them to alter their practices--it will not be sufficient merely to tell researchers not to use the statistical tests, or to use them more judiciously. (p. 26)

References

Abelson, R. P. (1997). A retrospective on the significance test of 1999 (If there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), What if there were no significance tests? (p. 117-141). Mahwah, NJ: Erlbaum.

American Psychological Association. (1994). Publication manual of the American Psychological Association (4th ed.). Washington, DC: Author.

Azar, B. (1997). APA task force urges a harder look at data. The APA Monitor, 28 (3), 26.

Carver, R. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Carver, R. (1993). The case against statistical significance testing, revisited. Journal of Experimental Education, 61, 287-292.

Cohen, J. (1969). Statistical power analyses for the behavioral sciences. New York: Academic Press.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

Cohen, J. (1994). The earth is round ($p<.05$). American Psychologist, 49, 9971003.

Cortina, J. M., & Dunlap, W. P. (1997). Logic and purpose of significance testing. Psychological Methods, 2, 161-172.

Cronbach, L. J. (1975). Beyond the two disciplines of psychology. American Psychologist, 30, 116-127.

Daniel, L. G. (in press). Statistical significance testing: A

historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. Research in the Schools.

Daniel, W. W. (1977). Statistical significance versus practical significance. Science Education, 61, 423-427.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. Psychological Methods, 1, 379-390.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5, 3-8.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. American Psychologist, 52, 15-24.

Hunter, J.E. (1997). Needed: A ban on the significance test. Psychological Science, 8(1), 3-7.

Keselman, H.J., Huberty, C.J, Lix, L.M., Olejnik, S., Cribbie, R., Donahue, B., Kowalchuk, R.K., Lowman, L.L., Petoskey, M.D., Keselman, J.C., & Levin, J.R. (in press). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses. Review of Educational Research.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. Educational and Psychological Measurement, 56, 5.

Kupfersmid, J. (1988). Improving what is published: A model in search of an editor. American Psychologist, 43, 635-642.

Levin, J.R. (1998). To test or not to test H_0? Educational and Psychological Measurement, 58, 311-331.

McLean, J. E., & Ernest, J. M. (in press). The role of statistical significance testing in educational research. Research in the

Schools.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Morrison, D. E., & Henkel, R. E. (Eds.). (1970). The significance test controversy. Chicago: Aldine.

Nelson, N., Rosenthal, R., & Rosnow, R.L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. American Psychologist, 41, 1299-1301.

Neyman, J., & Pearson (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Biometrika, 29A, Part I: 175-240; Part II: 263-294.

Nix, T. W. & Barnette, J. J. (in press). The data analysis dilemma: ban or abandon. A review of null hypothesis significance testing. Research in the Schools.

Oakes, M. (1986). Statistical inference: A commentary for the social and behavioral sciences. New York: Wiley.

Rindskopf, D. M. (1997). Testing 'small', and null, hypotheses: Classical and Bayesian approaches. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), What if there were no significance tests? (p. 407-490). Mahwah, NJ: Erlbaum.

Robinson, D. & Levin, J. (1997). Reflections on statistical and substantive significance, with a slice of replication. Educational Researcher, 26 (5), 21-26.

Rosenthal, R. & Gaito, J. (1963). The interpretation of level of significance by psychological researchers. Journal of Psychology, 55, 33-38.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 12176-1284.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. Psychological Bulletin, 57, 416-428.

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. Psychological Methods, 1(2), 115-129.

Schmidt, F.L., & Hunter, J.E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), What if there were no significance tests? (pp. 37-64). Mahwah, NJ: Erlbaum.

Shea, C. (1997). Psychologists debate accuracy of "significance test." Chronicle of Higher Education, 42 (49), A12, A16.

Thompson, B. (1993). Theme issue: Statistical significance testing in contemporary practice. Journal of Experimental Education, 61 (4).

Thompson, B. (1994). The concept of statistical significance testing (An ERIC/AE Clearinghouse Digest #EDO-TM-94-1). Measurement Update, 4(1), 5-6. (ERIC Document Reproduction Service No. ED 366 654)

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. Educational Researcher, 25 (2), 26-30.

Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. Educational Researcher,

26(5), 29-32.

Thompson, B. (1998a, April). Five methodology errors in educational research: The pantheon of statistical significance testing and other faux pas. Invited address presented at the annual meeting of the American Educational Research Association, San Diego. (ERIC Document Reproduction Service No. ED forthcoming)

Thompson, B. (1998b). In praise of brilliance: Where that praise really belongs. American Psychologist, 53, 799-800.

Thompson, B. (1998c). Review of What if there were no significance tests? by L. Harlow, S. Mulaik & J. Steiger (Eds.). Educational and Psychological Measurement, 58, 332-344.

Thompson, B. (in press-a). If statistical significance tests are broken/misused, what practices should supplement or replace them? Theory and Psychology. Invited address presented at the 1997 annual meeting of the American Psychological Association, Chicago).

Thompson, B. (in press-b). Journal editorial policies regarding statistical significance tests: Heat is to fire as p is to importance. Educational Psychology Review.

Thompson, B. (in press-c). Statistical significance tests, effect size reporting, and the vain pursuit of pseudo-objectivity. Theory and Psychology.

Thompson, B & Snyder, P. A. (1998). Statistical significance and reliability analyses in recent Journal of Counseling and Development research articles. Journal of Counseling and Development, 76, 436-441.

Vacha-Haase, T., & Nilsson, J.E. (1998). Statistical significance

reporting: Current trends and usages within *MECD*. Measurement and Evaluation in Counseling and Development, 31, 46-57.

Zuckerman, M., Hodgins, H.S., Zuckerman, A., & Rosenthal, R. (1993). Contemporary issues in the analysis of data: A survey of 551 psychologists. Psychological Science, 4, 49-53.

NHST: A Continuum of Views:

Keep it!	**Keep it, but use it cautiously**	**If you're going to use it, use it right!**	**Ban it!**
Abelson (1997)	Levin (1998)	(i.e. emphasize effect size and result replicability)	Carver (1978, 1990)
Cortina & Dunlap (1997)	Robinson & Levin (1997)	Cohen (1994)	Hunter (1997)
Frick (1996)	**Keep it, but look at effect sizes**	Thompson (1993, 1996, 1998a, 1998b, 1998c, in press-a, in press-b, in press-c)	Schmidt & Hunter (1997)
Hagen (1997)	Kirk (1996)		
Rindskopf (1997)			

Three Major Limitations of NHST Reported in the Literature:

1. p values cannot themselves be used as indices of effect size.

2. Unlikely results (i.e. those with a small p value) are not necessarily interesting or important.

3. p values do not inform us about result replicability.